

ABSTRACT OF THE DISCLOSURE

Disclosed is a method for customizable schema-guided conversion of plain-text documents, rich-text documents and textual data records to an XML-compatible structured form. The method makes substantial use of element content model definitions from a chosen target XML schema/DTD to optimize, closely guide, and disambiguate element pattern matching and recognition. Highly granular structure can be inferred, in best possible conformance with the schema. One embodiment operates based on a finite state machine derived via recursive aggregation of the schema element content models. Additionally disclosed is a method for automated document structuring within the environment of an XML-enabled wordprocessor application. The method entails using the host's API to perform element pattern search and matching and to apply markup to the document in accordance with the inferred XML structure. A GUI framework integrated in the wordprocessor workspace can be provided for developing and executing document conversion/structuring definitions.